Asia Pacific Journal of Information Systems
Vol. 26 No. 1 (MARCH 2016), 66-79

ISSN 2288-5404 (Print) / ISSN 2288-6818 (Online)
http://dx.doi.org/10.14329/apjis.2016.26.1.66

Purchase Prediction by Analyzing Users' Online Behaviors Using Machine Learning and Information Theory Approaches

Minsung Kim[a], Il Im[b,*], Sangman Han[C]

[a] *Manager, Big Data Solution Business Team, SK telecom, Korea*
[b] *Professor, Information systems at School of Business, Yonsei University, Korea*
[C] *Professor, Marketing at SKK Business School, Sungkyunkwan University, Korea*

A B S T R A C T

The availability of detailed data on customers' online behaviors and advances in big data analysis techniques enable us to predict consumer behaviors. In the past, researchers have built purchase prediction models by analyzing clickstream data; however, these clickstream-based prediction models have had several limitations. In this study, we propose a new method for purchase prediction that combines information theory with machine learning techniques. Clickstreams from 5,000 panel members and data on their purchases of electronics, fashion, and cosmetics products were analyzed. Clickstreams were summarized using the 'entropy' concept from information theory, while 'random forests' method was applied to build prediction models. The results show that prediction accuracy of this new method ranges from 0.56 to 0.83, which is a significant improvement over values for clickstream-based prediction models presented in the past. The results indicate further that consumers' information search behaviors differ significantly across product categories.

Keywords: Predictive Modeling, Information Theory, Machine Learning, Random Forests

Ⅰ. Introduction

The explosive increase in the volume of available data and advances in data processing methods have made big data analysis possible in recent years (Madden, 2012). Companies can now collect and analyze their customers' clickstreams and other data to personalize the services they provide, which can give them a critical competitive advantage in many industries. Personalized services commonly include coupons specific to users' interests (Lee et al., 2011) and dynamic re-organization of the information being displayed on users' screens such as recommendation lists (Im and Hars, 2007; Zhang et al., 2011).

The importance of behavioral prediction based

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2013S1A3A2055050)

* Corresponding Author. E-mail: il.im@yonsei.ac.kr Tel: 82221235480

on big data has increased along with the increase in available data (Ka and Kim, 2014). However, despite the numerous past studies on behavior prediction through clickstream analysis (Montgomery et al., 2004; Park et al., 2015; Senecal et al., 2014; Van den Poel and Buckinx, 2005), most of the studies used raw clickstream data to build their prediction models. In this study, we propose a new method to predict customers' purchasing behavior that effectively handles the vast amount of data resulting from clickstreams by applying the information theory and a machine learning technique. Data collected from 5,000 online panels for three months and the associated purchase data were analyzed to test the method.

Ⅱ. Consumers' Purchase Decision-Making Processes

Consumers' offline purchase decision-making processes generally consist of four phases – need recognition, information search, alternative evaluation, and final decision-making (Taylor, 1974). The main purpose of the information search phase is to reduce uncertainty. In circumstances of high uncertainty, consumers will continue to search for more information (Urbany et al., 1989). In the earlier stage of purchase decision-making, consumers consider a large number of alternatives (Fotheringham, 1988). Conversely, in the later stages, consumers reduce their consideration to three or four alternatives before they make a final decision (Awad et al., 2006; Manrai and Andrews, 1998). If consumers follow this same decision-making process online, analyzing their clickstream patterns should enable us to predict their propensity to purchase a given product. For example, if a consumer stops searching for information after doing so actively for some time, we may take it as an indication that he/she is entering the final decision-making phase.

2.1. Purchase Delay in Online Shopping

It is commonly known that consumers often delay making final decisions even after they have completed searching for information and evaluating alternatives (Greenleaf and Lehmann, 1995). The interval of delay varies depending on the reasons for delay. Most common reasons for delay are waiting for the possibility of better alternatives, the need for more information, and waiting for the price to drop or the quality to improve (Greenleaf and Lehmann, 1995).

The expansion of e-commerce in the years since the year 2000 has increased consumer choice, making it harder to make final purchase decisions (Schwartz and Kliban, 2004). One interesting consequence of this is that more and more consumers are choosing to defer their decision-making. In the online shopping context, it has been observed that many consumers do not press the 'check-out' button for a long time after they put products in their shopping carts; this may be an indication that they are searching for additional information (Cho et al., 2006).

2.2. Previous Studies on Purchase Prediction

In the past, researchers developed methods for predicting consumers' purchases by analyzing their clickstreams (Kim et al., 2004; Montgomery et al., 2004; Senecal et al., 2014; Van den Poel and Buckinx, 2005). Most of these studies used statistical analysis methods such as logit regression and showed that clickstream analysis is a good predictor of purchase behavior (Montgomery et al., 2004; Van den Poel and Buckinx, 2005).

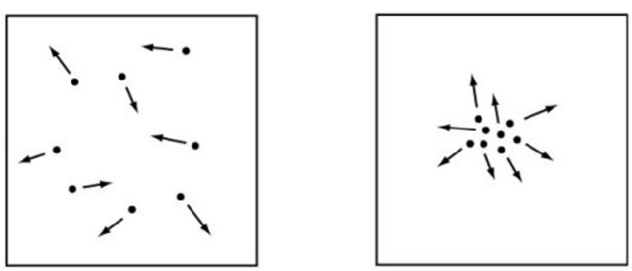
However, in these studies, raw clickstream data were used to create the prediction models. There are several limitations of using raw clickstream data as the input for behavior prediction. First, even after data cleansing through pre-processing, raw clickstream data are vast in volume and contain considerable amounts of noise. Because of the ubiquity of mobile devices, the volume of clickstream data is increasing exponentially every day. Therefore, analyzing all raw clickstream data in today's environment is not physically possible, nor does it give meaningful results. Second, using raw clickstream data, it is extremely hard, if not impossible, to model consumers' long-term behavior. In most previous studies, purchase behavior was modeled using a limited number of URLs. A single cycle of the process of making a purchase can take several weeks. Clickstream data collected in this long cycle may entail a large number of URLs (thousands, in some cases), which weakens the accuracy and feasibility of these prediction models. Third, with clickstream data, modeling visits to multiple sites is extremely difficult. In most previous studies, clickstream data were collected from a single site. In the information search phase, however, it is common for consumers to visit multiple sites, such as manufacturer's sites and consumer review sites, in order to obtain diverse information. Sites visited vary from person to person. Therefore, new methods that can incorporate a large volume of clickstream data and multiple sites must be developed.

2.3. Shannon's Information Theory

According to Shannon's information theory (Shannon, 1948), 'entropy' is the average amount of information contained in each message received. A 'message' here can be an event, a sample, or a character drawn from a distribution or data stream. <Figure 1> shows the concept of entropy as a measure of disorder. Higher entropy means a higher level of disorder, while lower entropy means a lower level of disorder.

In the context of information processing, entropy means the extent to which diverse information is included in the message. In a sense, entropy is similar in this context to 'uncertainty' or 'unpredictability.' Higher entropy means higher uncertainty or lower predictability. Conversely, lower entropy means lower uncertainty or higher predictability. Although the concept of entropy was coined in the natural science area, it has also been used in management research. For example, a study used entropy to measure dispersion of word of mouth across multiple newsgroups (Godes and Mayzlin, 2004).

When applied to consumers' purchase decision-making processes, Shannon's information theo-



<Figure 1> Entropy as a Measure of Disorder

ry implies that the level of entropy varies across different phases of the purchase process. Entropy would be higher in the early phase because consumers have higher uncertainty at this point and they need more diverse information. On the other hand, entropy would decrease in the later phase because consumers have already acquired sufficient information and uncertainty is lower. Information theory is in line with many studies of consumer behavior, which demonstrate that consumers conduct active information searches and compare a large number of product alternatives in the earlier shopping phase to reduce uncertainty, but compare only a few final alternatives and conduct limited information searches in later phases (Urbany et al., 1989).

From consumer research and studies based on information theory, we predict that consumer purchase behavior can be predicted by measuring entropy in the context of online information searching. In this study, we transform data collected from clickstreams to a measure of entropy and use it for finding patterns in online behavior and predicting purchases. Formula (1) shows the entropy of each user in the context of online behavior. The formula shows that entropy increases as users visit more diverse URLs or visit all URLs equally frequently. Conversely, entropy decreases as they visit fewer URLs or visit some URLs more than other URLs.

$$E(p) = -\sum_{i=1}^{n} p_i \log_2 p_i \quad (1)$$

where

n: the number of unique URLs that the current user visited

P_i: the probability that the i^{th} URL will be visited = (the number of visits to the i^{th} URL / the number of visits to all URLs)

Ⅲ. Prediction Techniques in Big Data Analysis

Big data analysis methods are not entirely new. Big data analysis, in a sense, is a combination of conventional data analysis methods and technologies for storing and processing massive amounts of data in real time (Boyd and Crawford, 2012). However, newly developed real-time analysis capabilities brought about many innovations, such as fraud detection and defect prediction. The two main goals of big data analysis are prediction and description (Behrooz, 2005). Prediction refers to the determination of unknown or future values from the data in a database, while description is the summary of general but interesting characteristics of a dataset in an understandable way.

In this study, we develop a classification method to predict the behavior of purchasers and non-purchasers. Classification is a method of grouping data into pre-defined classes (Behrooz, 2005). Classification enables us to explain current data and classes, thus providing us with a model to explain future data (Stork et al., 2001). There are several prediction techniques available through classification: Bayesian inference (Duda et al., 2001), neural network approaches (Wilson and Sharda, 1994), decision tree-based methods (Osei-Bryson and Ngwenyama, 2011; Quinlan, 1986), support vector machines (Ren et al., 2015), and genetic algorithm-based approaches (Lau et al., 2012; Punch III et al., 1993). These techniques are really just different kinds of supervised machine learning techniques. 'Random forests' is another machine learning technique that constructs a multitude of decision trees at training time and predicts the average or mode of the outputs of those decision trees (Breiman, 2001). The dataset used in this study for prediction is discrete

(i.e., daily entropy). Three different machine learning techniques – neural network analysis, supporting vector machine, and random forests – were considered. Since random forests method showed the highest prediction accuracy with a sample dataset, random forests was used for the prediction with the main data in this study.

Ⅳ. Empirical Analysis

In order to test the accuracy of the prediction method proposed above – the random forests technique applied to clickstreams – data collected from a panel between April 1 and June 30, 2014 were analyzed. The panel consisted of 5,000 volunteers and was managed by one of the largest advertising agencies in Korea. The company, which asked to remain anonymous, agreed to collaborate in this study of a prediction model. All clickstreams from panel members' PCs and smartphones (or other mobile devices) were collected by pre-installed agent software. Note that only clickstreams from web browsers, not from App usage activities, were collected. Since online tracking of financial transactions such as credit card payments is prohibited by law, actual purchase data were collected through a survey administered in July 2014. In the survey, panel members were asked to answer questions about their online information searching and buying behaviors. They were also asked to report whether they purchased products in any of three categories – electronics, fashion, and cosmetics – between June 11 and 20, 2014. This interval (11~20 June) was used instead of one specific date because many people did not recall the exact date of purchase.

4.1. Dataset

From the entire body of clickstream data, panel members were selected who reported in the survey that most of their product information searching was done online and purchased at least one product in the three categories above between June 11 and 20. Panel members who reported that they did not purchase any products in June were also selected as the control group. URLs from these purchasers and non-purchasers were used for building prediction models. Examples of clickstream data are shown in <Table 1>. The first column lists the time stamp (date and time) of the URLs, and the second column lists the unique user IDs of those who visited each URL.

4.2. Analytical Procedure

To facilitate data analysis, Cubrid was used for data storage/retrieval and R® ('randomForest' and

<Table 1> Clickstream Data

Date & Time	User ID	URL
2014-04-18 23:43:28	20121117B2E808BBDDFF	http://m.blog.daum.net/hjeom01/124
2014-04-18 23:45:23	20121117B2E808BBDDFF	http://m.blog.naver.com/halel1226/100205448197
2014-04-18 23:46:34	20121117B2E808BBDDFF	http://m.blog.naver.com/iteos/201541707
. . .	. . .	. . .

'reshape2' packages) was used for model building and evaluation. Some pre-processing of clickstreams was performed before the main analysis, as follows:

- Clickstreams were sorted by User ID and time.
- Clickstreams were grouped based on the purchased product (electronics, fashion, and cosmetics). If a user purchased two or more products, his/her clickstreams were copied into all corresponding groups.
- For each product category, clickstreams were re-arranged around the purchase date. As mentioned above, since the purchasers included in our sample bought products between June 11~20, we set June 15th as the purchase date for these people and then labeled clickstreams based on the time interval from the purchase date. For example, clickstreams collected on June 14th were labeled as d-1 (one day before purchase), those collected on June 13th as d-2 (two days before the purchase), and so forth.

After pre-processing, daily values for entropy for each user were calculated using formula (1). A sample final dataset after pre-processing is shown in <Table 2>. The data in the table show the entropies by user and date. 'Output' refers to whether or not the user purchased the product (1 = purchased, 2 = not purchased). X7 is entropy one day before purchase, X6 is entropy two days before the purchase, and so forth.

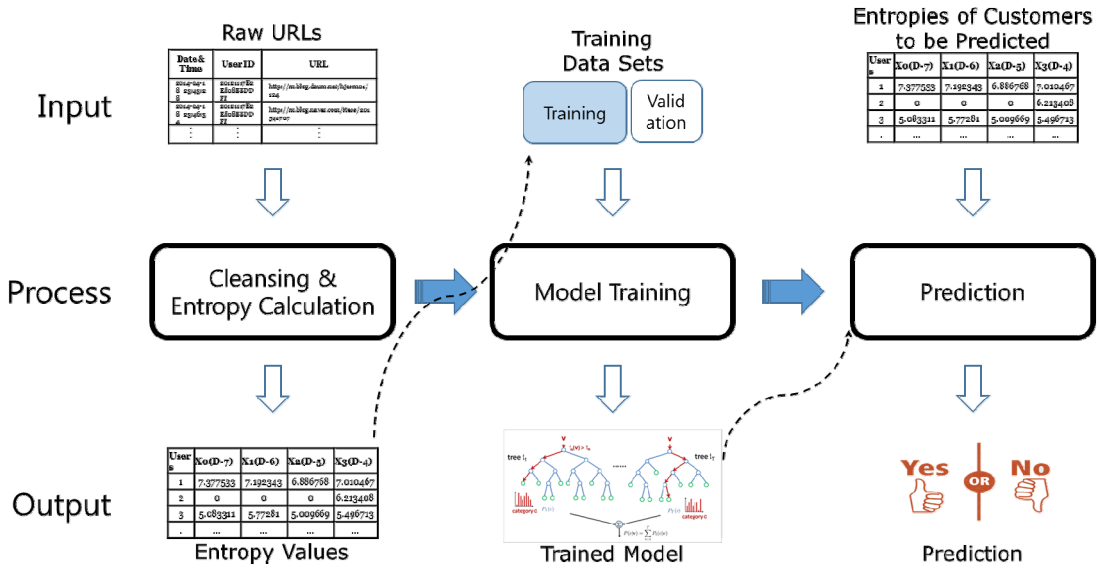
Adjustments were made for weekends and holidays. The search volume changed substantially depending on the day of the week and during holidays. Therefore, entropy values were adjusted to remove weekend and holiday effects as follows:

- For each user, average values for entropy for different days (Monday, Tuesday, etc.) were calculated.
- For each entropy value, the deviation between the entropy value and the average of the corresponding day was calculated. For example, if entropy for a certain Wednesday was 3.7, and the average value for entropy for all Wednesdays for a given user was 2.5, the final entropy value for that user in the analysis was determined to be 1.2 (= 3.7 – 2.5).
- Entropy values on holidays were replaced with the average entropy value for the corresponding day.

Then, entropy values for non-purchasers who reported that they did not purchase products in any of the three categories in June 2014 were calculated

<Table 2> Final Dataset after Pre-Processing

Users	X0(D-7)	X1(D-6)	X2(D-5)	X3(D-4)	X4(D-3)	X5(D-2)	X6(D-1)	X7(D-0)	Output
1	7.377533	7.192343	6.886768	7.010467	6.497301	7.157231	7.29089	6.676898	1
2	3.209045	4.106475	3.234869	6.32361	4.368932	6.896528	6.285387	7.362126	1
3	5.401263	6.192441	6.587059	6.565418	6.464482	5.732012	6.129243	6.082805	1
4	6.128106	4.840796	6.167554	6.215174	6.216096	6.000465	5.810154	6.991738	1
5	0	0	0	6.213408	5.948327	5.91606	4.637926	5.761245	1
6	5.083311	5.77281	5.009669	5.496713	5.556627	5.43734	5.659854	6.360638	0
7	0	0	0	6.314161	6.839494	7.247681	6.955742	6.864425	0
.	…	…	…	…	…	…	…	…	…



<Figure 2> Overall Prediction Procedure

<Table 3> Confusion Matrix for a Binary Classification Model

		Prediction	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

in the same way. The number of non-purchasers was 36, and they generated 2,512,862 URLs. The overall process of model training and prediction is shown in <Figure 2>.

The random forests method was applied to the final dataset of each product category. For each dataset, 500 decision trees were generated with 8 variables and 30 simulations were conducted using the 10-fold cross-validation method. With this method, the dataset is divided into 10 subsets, one of which is selected as the validation set, and the other 9 sets are used as the training sets. Since 10 iterations are run with different validation sets, all subsets are evaluated. The average value for accuracy of these results was used as the final value. Accuracy in a binary classification model for the purpose of purchase prediction, as used in this study of purchasers and non-purchasers, is defined as the ratio of true positive and true negative predictions to total predictions (Provost et al., 1998), as shown in <Table 3> and formula (2).

$$\text{Accuracy} = \frac{a + d}{a + b + c + d} \quad (2)$$

Ⅴ. Results

As described above, data from three product categories – electronics, fashion, and cosmetics – were analyzed. Different product categories had different

levels of accuracy. According to the literature, electronics products are generally considered the products of highest consumer involvement among these three categories (Gu et al., 2012). Fashion products generally elicit higher involvement than cosmetics (Laurent and Kaperer, 1985).

5.1. Electronic Products

In June 2014, 48 people in our sample purchased electronic products, generating 2,780,709 URLs. <Figure 3> shows the average entropy values of purchasers (solid line) of electronic products and non-purchasers (dotted line). A significant drop in entropy is evident among purchasers about 9 days before the date of purchase. No significant decline in entropy is shown for non-purchasers. This implies that for purchasers of electronics products in our sample, information search activity is substantially reduced about 9 days prior to actual purchase, in accordance with the purchase delay discussed above. However, purchasers later resume their search activities, gradually reducing them after D-3 until they make a purchase.

5.2. Fashion Products

In June 2014, 77 people in our sample purchased fashion products; these users generated 4,954,533 URLs. <Figure 4> shows the average values for entropy of purchasers (solid line) and non-purchasers (dotted line) of fashion products. For purchasers of fashion products, search activity declines about 4 days before the purchase date until the actual purchase is made. Compared to electronic products, the period of purchase delay seems shorter for fashion products (9 days vs. 4 days).

5.3. Cosmetics

In June 2014, 66 people purchased cosmetics and generated 15,956,025 URLs. <Figure 5> shows the average entropy values of purchasers (solid line) and non-purchasers (dotted line). No significant drop in entropy among purchasers of cosmetics is observed. This implies that purchasers in this category did not delay before purchasing, but instead made their purchases as soon as they decided to buy.

5.4. Prediction Accuracy

The numbers of purchasers and non-purchasers were not the same. Therefore, the bootstrapping method was employed to equalize the two groups in terms of size. The accuracy of the random forests method was compared with random prediction. The accuracy of random prediction was 0.5 because the numbers of purchasers and non-purchasers were the same. Since the accuracy measure in this study is a ratio (% of accurate prediction), a non-parametric tests appears to be more appropriate for testing accuracy differences than parametric tests. Since the prediction process was repeated 30 times for each product category, there are 30 prediction accuracies for each product category. For each comparison, a 2 x 2 (random vs. proposed method and correct vs. incorrect prediction) cross-table was created and a Chi-square test was conducted against the null hypothesis, "Accuracies of the proposed method are not different from those of random prediction method." The test results are summarized in <Table 4>.

For electronic products, all p-values for the 30 tests were greater than 0.05 while the p-values for fashion and cosmetic products were below 0.01 for all tests. Therefore, we conclude that our model per-

<Figure 3> Entropy of Purchasers and Non-Purchasers (Electronic Products)

<Figure 4> Entropy of Purchasers and Non-Purchasers (Fashion Products)

<Figure 5> Entropy of Purchasers and Non-Purchasers (Cosmetics)

<Table 4> Prediction Accuracy

Product	Mean	Random	Accuracy Improvement	Mean of Chi-square	Mean of p value
Electronic	0.56	0.5	12%	0.57	0.609
Fashion	0.83	0.5	66%	45.07	0.000
Cosmetic	0.73	0.5	46%	24.08	0.000

formed significantly better than the random model for fashion and cosmetic product categories.

Compared to previous studies employed conventional clickstream-based prediction models, our method shows better performance. For example, in a past study (Van den Poel and Buckinx, 2005) using data from an online wine vendor, compatible accuracy levels of 0.68 ~ 0.70 were reported. Involvement level of Champaign is compatible with that of dress (Laurent and Kapferer, 1985). Therefore, wine is com-

parable with the fashion products examined in this study, which yielded an accuracy of 0.83 using our model. Although this is not a statistical test, there seems to be a substantial difference between past results and our new method.

Ⅵ. Discussion

The data analysis in this study reveals that consumers exhibit significantly different patterns in searching for information depending on product type. For example, for high-involvement products such as electronics products, consumers' information searching behaviors differ considerably from their search for information about low-involvement products such as cosmetics. The most dramatic reduction in information searching can be seen in the middle of the purchase process of electronics product (about 9 days before purchase). This implies that consumers dedicate a significant portion of their search activities to searching for information about high-involvement products, which results in salient patterns in infor mation searching. This reduction may be related to the purchase delay mentioned above. On the other hand, no significant changes or patterns in information searching behaviors are evident for low-involvement products such as cosmetics.

In addition, the highest accuracy was found for fashion products, which are considered medium-involvement products, while the lowest accuracy was seen for electronics products. One possible reason for the low prediction accuracy regarding electronic products observed in this study is as follows. Since high-involvement products are generally expensive, consumers hesitate to take action regarding purchase of these products. Therefore, probably more consumers in the high-involvement product category hesitated at the later stage of purchase and eventually became non-purchasers, which made predictions less accurate. On the other hand, consumers purchasing low-involvement products such as cosmetics expended less effort on searching for information. In this category, isolating purchasers from non-purchasers is difficult, which may lower the prediction accuracy for cosmetics compared to that of fashion goods. Conversely, since fashion goods are medium-involvement products, information searching behaviors are clearer, and consumers in this category execute their decisions with less hesitation.

Ⅶ. Implications, Limitations, and Future Research

The results of this empirical study showed that the behaviors of online consumers can be predicted by analyzing the entropy of their clickstreams. There are several theoretical and practical contributions of this paper.

From a theoretical viewpoint, by combining the information theory and machine learning techniques, we propose a useful method for purchase prediction. This new method can be used for effective prediction of consumers' online behaviors, especially their purchase behaviors. This study also confirms the trends in information search behaviors in consumer purchasing and the 'purchase delay' phenomenon observed in previous studies. In addition, we have shown that the tendency to delay purchasing varies both in terms of the delay interval and its salience across product categories. This study also demonstrates that new measures of characterizing consumers' online behaviors, other than analysis of clickstream data, can significantly improve prediction accuracy. This

revelation opens the door to a new area of research in big data analysis – summarization of the vast volume of data to improve characterization of consumer behaviors. We believe that in future, researchers may develop various methods in related areas to improve big data analysis.

From a practical viewpoint, we provide a more efficient method of predicting consumers' behaviors, which can be used in formulating strategies for promotion and personalized services. This study also shows that consumers' information search patterns vary depending on the type of products for which they search; prediction accuracy may also vary accordingly. This implies that managers should be careful when applying purchase prediction methods to their products. Careful assessment of product characteristics is necessary to determine whether our prediction technique is suitable.

This study has several limitations and suggests some avenues for future study. First, we examined only three types of products. Therefore, the results need to be validated with a more diverse variety of products. Second, only one machine learning technique (random forests) was tested in this study. Other machine learning techniques could be examined to see if there are any improvements in accuracy. Also, the purchase date was not precisely measured, leading to the possibility of error. Further studies with accurate information on searching and purchase behaviors must be conducted. Although indirect links between information searching behaviors and purchase delay were observed, direct relations were not tested in this study. Therefore, new studies are needed to determine how purchase delay is represented in information search behavior.

<References>

[1] Awad, N. F., Jones, J. L., and Zhang, J. (2006). *Does search mater? Using online clickstream data to examine the relationship between online search and purchase behavior.* Paper presented at the Proceedings of the Twenty-Seventh International Conference on Information Systems.

[2] Behrooz, M.-B. (2005). *Data mining for a web-based educational system.* PhD thesis, Michigan State University, East Lansing, MI, USA, 2005. Adviser-Punch, William F.

[3] Boyd, D., and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662-679.

[4] Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

[5] Cho, C.-H., Kang, J., and Cheon, H. J. (2006). Online shopping hesitation. *CyberPsychology & Behavior, 9*(3), 261-274.

[6] Fotheringham, A. S. (1988). Note-Consumer Store Choice and Choice Set Definition. *Marketing Science, 7*(3), 299-310.

[7] Godes, D., and Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing Science, 23*(4), 545-560.

[8] Greenleaf, E. A., and Lehmann, D. R. (1995). Reasons for substantial delay in consumer decision making. *Journal of Consumer Research, 22*(2), 186-199.

[9] Gu, B., Park, J., and Konana, P. (2012). Research note-the impact of external word-of-mouth sources on retailer sales of high-involvement products. *Information Systems Research, 23*(1), 182-196.

[10] Im, I., and Hars, A. (2007). Does a one-size recommendation system fit all?: The effectiveness of collaborative filtering based recommendation systems across different domains and search modes.

ACM Transactions on Information Systems, 26(1), 1-30.

[11] Ka, H.-K., and Kim, J.-s. (2014). An empirical study on the influencing factors for big data intended adoption: Focusing on the strategic value recognition and TOE framework. *Asia Pacific Journal of Information Systems, 24*(4), 443-472.

[12] Kim, D.-H., Atluri, V., Bieber, M., Adam, N., and Yesha, Y. (2004, November 8 - 13). *A clickstream-based collaborative filtering personalization model: Towards a better performance.* Paper presented at the Proceedings of the 6th Annual ACM International Workshop on Web Information and Data Management, Washington, DC, USA.

[13] Lau, R. Y., Liao, S. S., Wong, K.-F., and Chiu, D. K. (2012). Web 2.0 environmental scanning and adaptive decision support for business mergers and acquisitions. *MIS Quarterly, 36*(4), 1239-1268.

[14] Laurent, G., and Kapferer, J.-N. (1985). Measuring consumer involvement profiles. *Journal of Marketing Research, 22*(1), 41-53.

[15] Lee, D.-J., Ahn, J.-H., and Bang, Y. (2011). Managing consumer privacy concerns in personalization: a strategic analysis of privacy protection. *MIS Quarterly, 35*(2), 423-444.

[16] Madden, S. (2012). From databases to big data. *IEEE Internet Computing, 16*(3), 4-6.

[17] Manrai, A. K., and Andrews, R. L. (1998). Two-stage discrete choice models for scanner panel data: An assessment of process and assumptions. *European Journal of Operational Research, 111*(2), 193-215.

[18] Montgomery, A. L., Li, S., Srinivasan, K., and Liechty, J. C. (2004). Modeling online browsing and path analysis using clickstream data. *Marketing Science, 23*(4), 579-595.

[19] Osei-Bryson, K. M., and Ngwenyama, O. (2011). Using decision tree modelling to support Peircian abduction in IS research: a systematic approach for generating and evaluating hypotheses for systematic theory development. *Information Systems Journal, 21*(5), 407-440.

[20] Park, J., Chung, Y., and Cho, Y. (2015). Using the hierarchical linear model to forecast movie box-office performance: The effect of online word of mouth. *Asia Pacific Journal of Information Systems, 25*(3), 563-578.

[21] Provost, F. J., Fawcett, T., and Kohavi, R. (1998). *The case against accuracy estimation for comparing induction algorithms.* Paper presented at the ICML.

[22] Punch III, W. F., Goodman, E. D., Pei, M., Chia-Shun, L., Hovland, P. D., and Enbody, R. J. (1993). *Further Research on Feature Selection and Classification Using Genetic Algorithms.* Paper presented at the ICGA.

[23] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning, 1*(1), 81-106.

[24] Ren, G., Hong, T., and Park, Y. (2015). Mutli-class SVM+MTL for the prediction of corporate credit rating with structured data. *Asia Pacific Journal of Information Systems, 25*(3), 579-596.

[25] Schwartz, B., and Kliban, K. (2004). *The paradox of choice: Why more is less*. New York: Ecco.

[26] Senecal, S., Kalczynski, P. J., and Fredette, M. (2014). Dynamic identification of anonymous consumers' visit goals using clickstream. *International Journal of Electronic Business, 11*(3), 220-233.

[27] Shannon, C. E. (1948). A note on the concept of entropy. *Bell System Tech. J, 27*, 379-423.

[28] Stork, D. G., Duda, R. O., Hart, P. E., and Stork, D. (2001). *Pattern Classification* (2nd ed.). New York: John Wiley & Sons.

[29] Taylor, J. W. (1974). The role of risk in consumer behavior. *Journal of Marketing, 38*, 54-60.

[30] Urbany, J. E., Dickson, P. R., and Wilkie, W. L. (1989). Buyer uncertainty and information search. *Journal of Consumer Research, 16*(2), 208-215.

[31] [Van den Poel, D., and Buckinx, W. (2005). Predicting online-purchasing behaviour. *European Journal of Operational Research, 166*(2), 557-575. doi:http://dx.doi.org/10.1016/j.ejor.2004.04.022

[32] Wilson, R. L., and Sharda, R. (1994). Bankruptcy prediction using neural networks. *Decision Support Systems, 11*(5), 545-557.

[33] Zhang, T., Agarwal, R., and Lucas Jr, H. C. (2011).

The value of IT-enabled retailer learning: personalized product recommendations and customer store loyalty in electronic markets. *MIS Quarterly, 35*(4), 859-882.

◆ About the Authors ◆

Minsung Kim

Minsung Kim currently works for SK Telecom. She received her master's degree from School of Business, Yonsei University. She has worked for several companies including LG Electronics, MC Research Institute, and Korea Development Institute (KDI). Her research interests are recommendation systems, social network analysis, and artificial intelligence (AI).

Il Im

Il Im is a full professor of information systems at School of Business, Yonsei University. He received his Ph.D. from Marshall School of Business, University of Southern California. Prior to joining Yonsei University, he was an assistant professor in the Information Systems Department at New Jersey Institute of Technology (NJIT). His current research focuses on personalization technologies and their effects, the effects of social network systems, and technology acceptance.

Sangman Han

Sangman Han is a full professor of marketing at SKK Business School, Sungkyunkwan University. He received his Ph.D. degree from Columbia University and he has been at Hong Kong University of Science & Technology as an assistant professor. His current research areas are marketing analytics, social network marketing and consumer decision journey.

Submitted: September 8, 2015; 1st Revision: December 8, 2015; Accepted: December 29, 2015